Predicting College Students Dropout using EDM Techniques

Anjana Pradeep
P.G. Scholar
Department of Computer Science and Engineering
St. Josephs College of Engineering and Technology
Kerala, India

Jeena Thomas
Assistant Professor
Department of Computer Science and Engineering
St. Josephs College of Engineering and Technology
Kerala, India

ABSTRACT

This study examines the factors affecting students' academic performance that contribute to the prediction of their failure and dropout using educational data mining techniques. This paper suggests the use of various classification techniques to identify the weak students who are likely to perform poorly in their academics. WEKA, an open source data mining tool was used to evaluate the attributes predicting student failure. The data set is comprised of 67 attributes of 150 students who have enrolled in B. Tech Degree Course registered for the academic year 2014-18 in a reputed college in Kerala affiliated to M.G University, Kerala, India. Various classification techniques like induction rules and decision tree have been applied to the data. The results of each of these approaches have been compared to select the one that achieves high accuracy.

Keywords

Educational Data Mining (EDM); Classification; Dropout; WEKA.

1. INTRODUCTION

There has been a growing interest and concern about the problem of students failure and determining the main factors contributing to this problem in recent years. Research and practical experimentation at colleges and universities across the country are revealing promising solutions that could enable colleges and universities to increase graduation rates while maintaining or reducing costs and ensuring that all students receive a high-quality educational experience that is tailored to their needs, academic abilities, and career or employment goals[1].

The number of students who are dropping from colleges out each year has been increasing. Therefore, many researches were aimed at examining the factors that affect the low performance of students at different educational levels like primary, secondary and higher [2][7]. The amount of information stored in educational databases is rapidly increasing due to the advancement in the field of information technology. These databases contain a wealth of data about students and are a gold mine of valuable information. The difficult task here is to identify and classify the valuable information hidden in those databases[4]. A very promising solution for this problem is the use of knowledge discovery in databases or data mining in education called Educational Data Mining (EDM)[3].

Data Mining is a non-trivial process of identifying valid, interesting, novel, useful, and ultimately understandable patterns hidden in data. Educational Data Mining is a field that exploits machine-learning, statistical and data mining algorithms over various types of educational data to resolve educational research issues[6]. EDM focuses on developing methods to explore unique patterns of data and to better understand students and settings in which they learn[3]. This paper aims at predicting college students' failure using the techniques of data mining. The factors that most influence failure in young students are to be detected using classification techniques. As the data have high dimensionality and highly unbalanced the use of different approaches are proposed. Several experiments are performed in order to obtain the highest classification accuracy. Several experiments were performed in order to obtain the highest classification accuracy. In a first experiment,8 classification algorithms were executed using all available information (67 attributes). In a second experiment, only the best attributes selected (13) were used. In a third experiment, the executions were repeated by using re-balanced data files. The outcomes have been compared and the models with the best results are shown. A case study is also done using the model constructed with the help of classification technique to predict the results of students.

2. RELATED WORKS

Most cited literature survey papers in Educational Data Mining have been by Romero and Ventura [5], Ryan Baker [15], and Romero and Ventura [6] which indicate performance prediction as one of the emerging field of educational data mining. Various subject performance attributes have been used by Paris, Affecndy and Musthafa[16] to predict final CGPA(Cumulative Grade Point Analysis) of Bachelor of Computer Science students of a Malaysian University. Various Bayesians Classification techniques have been used and a comparative study suggests that Ensemble method gives best overall accuracy.

Two diverse populations, Can Thao University of Vietnam and Asian Institute of Technology were considered by Paul et al. and achieved similar levels of accuracy of prediction performance for both the population. Cheewaprakobkit[18] considered 22 attributes of 1600 students records of Thailand University registered between the academic year 2001 and 2011 and decision tree algorithms and neural network algorithm were applied to most important factors affecting students' academic achievement. Decision tree proves to be a better classifier than the neural network with 1.31% more accuracy. Number of hours worked per semester, additional English course, number of credits enrolled per semester and marital status of the students are major factors affecting students' performance.

The importance of 24 predictor variables including demography, scores in maths, Turkish, religion and ethics, science and technology and level determination exams etc. have been ranked for predicting Turkish secondary education placement result [19]. Application of Artificial Neural Network, Support Vector Machine, Multiple Regression and Decision indicated that most important predictor variables are determination exam, scholarship, number of siblings, and success level in Turkish Language etc. Few personality traits like motivation of study, interests, learning environment , along with demographic details and previous academic performance have been considered by Wook et al.[20] to predict CGPA of Computer Science graduate and finally to find out students who are at risk of failing .

Tree classifiers as well as non-tree classifiers have been applied by Bidgoli et al.[21] to predict the grades of students enrolled with online education Latest Learning Online Network with Computer Assisted Personalized Approach (LON – CA PA) developed at Michigan State University. It was found that the prediction accuracy was enhanced with the use of combination of multiple classifiers.

The students' attitude towards study and scores earned at high school has been taken as attributes to predict the grade of first year students. The data for the model was collected through a questionnaire survey conducted during the summer semester at the Faculty of Economics in Tuzla. The model of students' success is measured with the success in the course "Business Informatics". Score of entrance exam, study material and average weekly hours devoted to studying have been found to have maximum impact while number of household member distance of residence and gender have been found to have least impact. Naive Bayes is found to be better classifier than J48[22].

Chi-squared Automatic Interaction Detector (CHAID) has been used by Ramaswami and Bhaskaran [23] to classify XII grade students of selected Tamil Nadu schools. Apart from demographic details, students' health, tuition, care of study at home etc. have been studied. Prediction Accuracy obtained was 44.69, and potential influences variables were found to be X grade marks, location of school, private tuition etc.

Various decisions tree algorithms like C 45, Random Forest, BF Tree, RepTree and Functions like logistic RBF Network, Rule Induction algorithms like JRip and Naive Bayes were used by Shah [24] to categorize students of BBA program enrolled at University of Karachi, Pakistan. Out of 42 independent variables 5 best variables having highest effect in determining performance is considered. Random Forest decision tree algorithm has proven to be the most accurate classifier than J48 decision tree, BF Tree, Rep Tree and JRip rule.

Data mining classification techniques has been applied on 10330 students by Kabakchieva[25] with 14 attributes including personal profile, secondary educational score, entrance exam score, admission year etc. The students are classified into five categories *Excellent, Very Good, Good, Average and Bad*. 10 fold cross validation and percentage split is used for all the classifiers like J48, Bayesian, K-nearest Neighbour, OneR and JRip. J48 has found to be most suitable of all classifiers.

The data of 346 first year students of an Engineering college was used by Kabra and Bichkar[26] to predict whether a student will PASS/FAIL or get promoted(When he fails in 3 theory and 2 practical subjects). Their demographic data (category, gender etc.), past performance data (SSC or 10th marks, HSC or 10,+ 2 exam marks etc.), address and contact number have been collected and used for the experiment. J48 algorithm in WEKA produces a prediction model with accuracy 60.46 %. The most important attribute in predicting student's performance is found to be HSCCET. The social attributes like category, parents' occupation, living location and other attributes like gender, medium at secondary level are found to be less relevant.

The students dropping out of an open polytechnic of New Zealand due to failure has been explored by Kovacic[27]. Enrolment data consisting of socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block), of 435 polytechnic students of Information System course were collected. The final label consisted of two categories PASS (those who completed the course) and FAIL (those who did not complete) were considered. Feature selection algorithms indicated that most important attributes for prediction are ethnicity, course programme and course block.

3. PROPOSED METHOD

The method proposed in this study for predicting the factors affecting the failure of students belongs to the process of Knowledge Discovery from Databases and Data Mining. The main steps in this method are:

1. Data gathering: Data may be obtained from many different and heterogeneous data sources. This stage comprises of gathering all available information on students. The set of factors that can affect the students' performance is first identified and collected from various sources of data available. This is then integrated into a single data set.

2. Data pre-processing: At this stage, the preparation of data set to apply the data mining techniques is done. Traditional pre-processing methods like data cleaning, data partitioning and data transformation of variables have to be applied. Due to problems of high dimensionality and imbalanced data, here we have also applied selection of attributes and re-balancing of data.

3. Data mining: DM algorithms are applied to analyse the factors affecting failure like a classification problem where a model is constructed. We propose the use of various classification algorithms and techniques that easily generates interpretable models like decision trees and induction rules. Finally these algorithms have been executed, examined, evaluated and compared in order to determine which one obtains the best result with high accuracy.

4. Interpretation: The obtained models are analysed to detect the problem of student failure in this stage. To achieve this, we interpret the factors that contribute more to the problem and how they are related are considered and assessed.

Next, a case study with data collected from Indian students is described in order to show the utility of the proposed method.

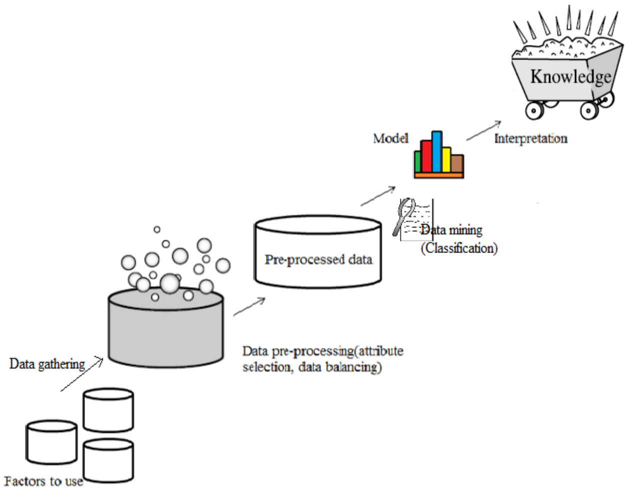

Figure 1: method used to predict student failure

3.1 Data Gathering

Due to the large amount of risk factors that contribute to the problem of failure of students it is also known as "one thousand factor problem". Some of the factors or characteristics that influence failure include cultural, social, family or educational background, demographics, socioeconomic status, psychological profile and academic progress. The information used in this study was collected from college students enrolled in Bachelor of Technology (B. Tech) programme at a reputed college in Kerala affiliated to Mahatma Gandhi University(MG University) for the 2013-18 academic year. B. Tech course offers a four-year education program that provides the student with scientific knowledge to continue studying the subjects in which they are interested. There are many specializations for B. Tech degree programme viz. Computer Science and Engineering(CSE), Civil Engineering(CE), Mechanical Engineering(ME),Electronics and Communication Engineering(ECE), Electrical and Electronics Engineering(EEE), Applied Electronics and Instrumentation Engineering(AEI), etc.

For this study, data about first-year B. Tech students of CSE and ECE have been used where most students are between the ages of 18 and 19, as this is the year when most of the students experience a new environment and infrastructure of study. There are 9 subjects to study during the first year of the course and these are common for students of all branches. The students will be considered as 'Fail' if they failed in at least one subject. All the information used in this study has been gathered from two different sources during the period from November, 2013 to April, 2015:

1) A specific and general survey was designed with the help of a questionnaire and administered to all students in the middle of the course. Its purpose was to obtain personal and family information to identify some important factors that could affect performance of students. It also provided a way to obtain the scores attained by students in 12th and 10th grades, Entrance Rank etc.

2) The Computer Science and Engineering Department and Electronics and Communication Department of the college provided the scores obtained by students in all the subjects in the first sessional examination conducted during the middle of the course.

Table 1. Attributes used and information sources

Source	Variable
General Survey	Name, Branch, Age, Sex, Religion, Caste, number of friends, number of hours spent studying daily, methods of study used, place normally used for studying, resources for study, study habits, parental encouragement for study, type of personality, having a physical disability, suffering a critical illness, regular consumption of alcohol, smoking habits, family income level, having a scholarship, living with ones' parent, mothers level of education, mothers occupation, fathers level of education, fathers occupation, number of brothers/sisters, position as the oldest/middle/youngest child, transport method used to go to college, distance to college, level of attentiveness during classes, level of boredom during classes, reasons for joining this course, difficulty level in EM1, difficulty level in EP, difficulty level in EC, difficulty level in EM, difficulty level in EG, difficulty level in BCE, difficulty level in BME, difficulty level in BLE, difficulty level in BEE&IT, level of motivation, taking notes in class, methods of teaching, too heavy a demand of homework, quality of college infrastructure, having a personal tutor, level of teachers concern for the welfare of each student, syllabus followed in Grade XII, Percentage in 12 and 10th grade, exercise habits, Score obtained in Mathematics in 12th grade, Score obtained in Physics in 12th grade, Score obtained in Chemistry in 12th grade, Quota of Admission, Entrance Rank.
CSE & ECE DEPARTMENT	Score in Engg. Mathematics(EM1), Engg. Physics(EP), Engg. Chemistry(EC), Engg. Mechanics(EM), Engg. Graphics(EG), Basic Civil Engg.(BCE), Basic Mechanical Engg.(BME), Basic Electrical Engg.(BLE), Basic Electronics Engg. and Information Technology(BEE&IT), Total.

3.2 Data Pre-Processing

Pre-processing of data is considered as a very important task in this work as we need quality and reliability of available information which directly affects the results attained. Before applying the data mining algorithms it is essential to carry out some pre- processing tasks such as data cleaning, integration, transformation and discretization. Pre-processing task includes finding incorrect or missing data. Erroneous data or ambiguous data may be corrected or removed, whereas missing data must be supplied. Pre-processing also includes removal of noise or outliers and collecting necessary information to model or account for noise. Transformation is the process of converting the data into a common format for processing. Some data may be encoded into more usable format.

Some specific pre-processing tasks were applied to the data set previously described so that data mining algorithms can be applied correctly. Firstly, students without 100% complete information were removed when all available data were integrated into a single dataset. All students who were unable to provide answers for our survey were excluded. Attributes that does not affect the classification results are removed. Name, Religion and Caste attributes have been removed as it does not have any relevance in classifying a student.

For providing a more comprehensible and compact view of data, continuous variables were transformed into discrete variables. For example, the numerical values of marks or scores obtained by students in each subject were changed to categorical values in the following way:

Table 2. Data transformation for marks obtained

Category	Marks obtained
Excellent	Between 95-100
Very good	Between 85-94
Good	Between 75-84
Regular	Between 65-74
Sufficient	Between 60-64
Poor	Between 40-59
Very Poor	Less than 40
Not Presented	-

The difficulty level in each subject had been collected by asking the students to rate their difficulty in each subject from a scale of 1-10. This information was transformed in the following way:

Table 3. Data transformation for difficulty level

Category	Rated Values
Easy	Between 0-2
Manageable	Between 3-5
Difficult	Between 6-7
Very Difficult	Between 8-10

Finally all this information was integrated in a single dataset and was saved in the .ARFF format of WEKA. The entire dataset was divided randomly into 10 pairs of training and test data files i.e., stratified tenfold cross-validation can be used to evaluate the classification algorithms. So after pre-processing our data set consist of 150 student records with 67 attributes.

There are two typical problems in the dataset that generally appears in these types of educational data.

1. High Dimensionality problem: Our dataset is highly dimensional, i.e., the number of attributes or features that contribute to the problem of failure is high. When there are a large number of attributes, some of them may not be meaningful for classification and it is likely that some of the attributes may be correlated.
2. Imbalanced Data problem: Here, we are classifying students with respect to their academic status, as PASS or FAIL. Majority of the students failed (98) and only minority passed (52). The problem with this imbalance in data is that learning algorithms tend to overlook less frequent classes and only pays attention to the most frequent class. As a result, the classifier may not be able to classify the data instances correctly.

To solve our first problem and to identify the attributes or features that have the greatest effect on our output variable, we carry out a feature selection algorithm. A wide range of attribute selection algorithms are available in WEKA that can be grouped in many ways. The way in which attribute can be evaluated is one of the most popular categorization of the algorithms and this way they can be grouped as filters and wrappers. Filters select and assess features independently of the learning algorithms and wrappers use the performance of learning algorithms to determine the desirability of an attribute subset.

WEKA provides several feature selection algorithms from which we have selected the following eight: CfsSubsetEval, ChiSquaredAttributeEval,FilteredAttributeEval, OneR Attribute Eval, Filtered SubsetEval, Gain RatioAttributeEval, InfoGain-AttributeEval, Relief FAttribute Eval. Table IV shows the results of applying these 8 algorithms of feature selection. The results obtained were ranked by these 8 algorithms to select the best attributes from our 67 available attributes. To find the ranking of the attributes, the number of times each attribute was selected by one of the algorithms was counted.

Table 4. Best selected attributes

Algorithm	Attribute Selected
CfsSubsetEval	level of attentiveness during classes, level of boredom during classes, difficulty level in EM1, taking notes in class ,Parental Encouragement for study, Score obtained in Mathematics in 12th grade ,Score in EC, EM, EG, BME,BLE, BEE&IT, Total
ChiSquared-AttributeEval	level of boredom during classes, difficulty level in EM1, Score obtained in Mathematics in 12th grade ,Score in EM1, EC, EM, EG, BME,BLE, BEE&IT, Total
Filtered-AttributeEval	level of boredom during classes, difficulty level in EM1, Family Income, Score obtained in Mathematics in 12th grade, Score in EM1, EC, EM, EG, BME,BLE, BEE&IT, Total
FilteredSubsetEval	Score obtained in Mathematics in 12th grade ,Score in EG, BME,BLE, BEE&IT, Total
GainRatio-AttributeEval	level of boredom during classes, difficulty level in EM1, Score obtained in Mathematics in 12th grade , taking notes in class , Score in EM1,EC, EM, EG, BME,BLE, BEE&IT, Total
InfoGain-AttributeEval	level of boredom during classes, Fathers Occupation, difficulty level

	in EM1, Fathers Education, Score in EM1,EC, EG, BME,BLE, BEE&IT, Total
OneRAttributeEval	Score in Mathematics; Score in Physics; Score in Chemistry; Score in Writing and reading; Score in English ; Score in Computer Science; Level of motivation.
ReliefFAttributeEval	Score in Mathematics; Score in Physics; Score in Chemistry; Score in Writing and reading; Score in English ; Score in Computer Science; Level of motivation; Age; percentage of marks obtained in 10th grade; Smoking habits; average score in ADMSN-EXM.

Table V shows the frequency of each attribute. From this table only those attributes appearing more than twice in these algorithms have been considered. Finally, we selected only the attributes with frequency greater than or equal to two (attributes selected by at least two algorithms). In this way, we can reduce the dimensionality of our dataset from the original 67 attributes to only the best 13 attributes.

Table 5. Most influencial attributes

Attribute	Frequency
Score in BME	8
Score in BEE&IT	8
Score in EG	8
Total	8
Score in BLE	7
Score in EC	7
Difficulty level in EM1	7
Score in EM1	6
Score in EM	6
Score obtained in Mathematics in 12th grade	6
level of boredom during classes	6
level of attentiveness during classes	2
Taking notes from class	2

It was also mentioned that the data set is imbalanced which happens when the number of instances in one class is much smaller than the number of instances in another class. To solve this problem care must be taken during the pre-processing stage itself by carrying out sampling or balancing of data. There are many data balancing or rebalancing algorithms that are commonly used and available in WEKA such as Synthetic Minority Over-Sampling Technique(SMOTE) algorithm[8]. In SMOTE algorithm, the minority class is over-sampled. This is done by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the 'k' minority class nearest neighbors. Neighbors are randomly chosen depending upon the amount of over-sampling required. Synthetic samples are generated by taking the difference between the feature vector under consideration and its nearest neighbor. Then we multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration which causes the selection of a random point along the line segment between two specific features. This method forces the decision region of the minority class to become more general in an effective way.

In this case, only the training set was balanced using SMOTE algorithm, and obtained 50% PASS students and 50% FAIL students without rebalancing the test files. The following tenfold cross validation files were obtained after performing all the previous procedures of data pre-processing:

1) Ten training and test files with all attributes (67).
2) Ten training and test files with only the best attributes (13).
3) Ten training and test files with only the best attributes (13); the training files are rebalanced using SMOTE.

3.3 Data Mining and Experimentation

This section describes the experiments and data mining techniques used for obtaining the prediction models of students' academic status at the end of the semester. Several experiments were performed in order to try to obtain the highest classification accuracy. In a first experiment, 8 classification algorithms using all available information (67 attributes)were executed. In a second experiment, only the best attributes selected (13) were used. In a third experiment, the executions of the algorithms were repeated by using re-balanced data files.

Classification that maps the data into predefined groups and classes is one of the most widely used data mining task. It is also called supervised learning which consists of two steps:

1. Model construction: Each tuple /sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set. The model can be represented as classification rules, decision trees, or mathematical formulae.

2. Model usage: This model constructed from the previous step can be used for classifying future or unknown objects. The known label of test sample is compared with the classified result. Accuracy rate can be defined as the percentage of test set samples that are correctly classified by the model.

In this paper, decision trees and rules induction algorithms are used as they are "white box" classification techniques; that is, they provide an explanation for the classification result and can be used directly for decision making.

A decision tree is a set of conditions organized in a hierarchical structure[12]. An instance is classified by following the path of satisfied conditions from the root of the tree until a leaf is reached, which will correspond with a class label. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. The decision tree is efficient and is thus suitable for large/small data sets. They are perhaps the most successful exploratory method for uncovering deviant data structure.

Rule induction algorithms usually employ a specific-to-general approach, in which obtained rules are generalized (or specialized) until a satisfactory description of each class is obtained[11]. Rule induction methods generalise the training set into rules that they can evaluate directly to classify new examples. These rules may be represented in many ways, including decision trees and modular rules. Rule induction systems evaluate the features of the training set and decide which ones to use to discriminate between the different classes.

Eight commonly used classical classification algorithms that are available in the well-known WEKA which is a DM software that have been used here[11][9]:

1) Four rule induction algorithms: JRip, which is a propositional rule learner; NNge, which is a nearest neighborlike algorithm; OneR, which uses the minimum-error attribute for class prediction; and Ridor, which is an implementation of the Ripple-Down Rule learner.

2) Four decision tree algorithms: J48 , which is an algorithm for generating apruned or unpruned C4.5 decision tree[10]; ADTree, which is an alternating decision tree; RandomTree, which considers K randomly chosen attributes at each node of the tree; and REPTree, which is a fast decision tree learner.

A decision tree can be directly transformed into a set of IF-THEN rules (which are obtained by rule induction algorithms), which are one of the most popular forms of knowledge representation due to their simplicity and comprehensibility. In this way a non-expert user of DM such a as teacher or instructor can directly use the output obtained by these algorithms to detect students with problems (classified as Fail) and to make decisions about how to help them and prevent their possible failure.

In the first experiment, all the classification algorithms were executed using tenfold cross-validation and all the available information, that is, the original data file with 67 attributes of 150 students. The results with the test files (an average of 8 executions) of classification algorithms are shown in Table VI. This table shows the rates or percentages of correct classifications for each of the two classes: Pass (TP rate) and Fail (TN rate) and the overall Accuracy rate (Acc). It can be seen in Table V that the percentage of accuracy obtained for total accuracy (Acc) and for Fail (TN rate) are high, but not for Pass(TP rate). Specifically, the algorithms that obtain the maximum values are: NNge (TN rate) and ADTree (TP rate and Acc).

Table 6. Result of classification using all attributes

	Algorithms	**TP Rate**	**TN Rate**	**Accuracy**
Rule Induction Algorithms	JRip	78.8	88.8	85.33
	NNge	61.5	**93.9**	82.67
	OneR	**80.8**	90.8	**87.33**
	Ridor	76.9	90.8	86
Decision Tree Algorithms	**ADTree**	**90.4**	**92.9**	**92**
	J48	76.9	91.8	86.67
	Random Tree	80.8	81.6	81.33
	REP Tree	76.9	90.8	86

In the second experiment, all the classification algorithms using tenfold cross-validation were executed and reduced the dataset (with only the best 13 attributes). These are attributes that contribute more to the classification results and are identified using the algorithms provided in WEKA tool. Table VII shows the results with the test files (the average of 8 executions) using only the best 13 attributes.

Table 7. Result of classification using the best attributes

	Algorithm	**TP Rate**	**TN Rate**	**Accuracy**
Rule Induction Algorithms	JRip	73.1	88.8	83.34
	NNge	**84.6**	89.8	**88**
	OneR	80.8	**90.8**	87.34
	Ridor	75	85.7	82
Decision Tree Algorithms	**ADTree**	**88.5**	**93.9**	**92**
	J48	76.9	**93.9**	88
	Random Tree	73.1	86.7	82
	REP Tree	78.88	87.8	84.67

When comparing the results obtained with the previous results obtained using all the attributes, that is, Table VI versus Table VII, it can be seen that in general all the decision tree algorithms have improved in measures like TN rate. Furthermore, with regard to the others measures there are some algorithms that obtain a slightly worse or slightly better value, but they are very similar in general to the previous ones. In fact, the maximum values obtained are now better than the previous ones obtained using all attributes. Again the algorithms that obtain these maximum values are ADTree (TP rate, TN Rate and Accuracy) and J48(TN Rate).As it can be seen from Tables VI and VII, a good classification of the minority class (Pass) have not been obtained yet. And this can be due to the fact that our data are imbalanced. This feature of the data is not desirable because it affects negatively in the results obtained. A classification algorithm tends to focus on classifying the majority class in order to obtain a good classification rate, but tends to forget the minority class.

In the third experiment, all the classification algorithms were again executed using tenfold cross-validation and the rebalanced training files (using The SMOTE algorithm) with only the best 13 attributes. The results obtained after re-executing the 8 classification algorithms using tenfold cross-validation are summarized in Table VIII. If we analyse and compare this table with the previous VII and VIII, we can observe that over half of the algorithms have increased the values obtained in all the evaluation measures, and some of them also obtain the new best maximum values in almost all measures except accuracy. The algorithm that have obtained the best results is ADTree again. The screen shots have been given in the appendix of the report.

Table 8. Classification result using data balancing

	Algorithm	**TP Rate**	**TN Rate**	**Accuracy**
Rule Induction Algorithms	**JRip**	**99**	**96.9**	**98.02**
	NNge	97.1	**96.9**	97.02
	OneR	87.5	91.8	89.60
	Ridor	96.2	96.9	96.53
Decision Tree Algorithms	**ADTree**	**100**	**99**	**99.5**
	J48	96.2	92.9	94.55
	Random Tree	97.1	96.9	97.02
	REP Tree	95.2	94.9	95.05

3.4 Interpretation of Results

In this section, some of the rules discovered by the algorithms that obtained maximum values for the evaluation measures are examined. Comparison of their interpretability and usefulness for identifying the risk of student failure is discussed. Decisions can be made about how to help students as early as possible is also done. These rules present some relevant factors and relationships among these factors that lead a student to pass or fail. The rules of algorithms that showed maximum TP Rate (Pass) , TN Rate (Fail) and Accuracy are shown in the following tables.

In the model shown in Table IX it is observed that the algorithm JRip discovers few rules. With respect to the attributes that are associated to Fail, they are mostly concerning marks, indicating that the student failed if they obtained Very Poor (i.e. less than 40 marks) score in Basic Electrical Engineering(BLE), Basic Electronics and Information Technology(BEE&IT), Basic Mechanical Engineering(BME) and Engineering Graphics(EG). There are other attributes like total marks which indicate that students who could not score more than 482 marks tend to fail.

Table 9. Rules generated by JRIP algorithm using the best 13 attributes and data balancing

```
(Total <= 577) and (Total <= 482) => Result=F (60.0/1.0)
(Total <= 596) and (BLE = Very Poor) => Result=F
(15.0/0.0)
(Total <= 595) and (BEE&IT = Very Poor) => Result=F
(8.0/0.0)
(BME = Very Poor) => Result=F (7.0/0.0)
(BEE&IT = Very Poor) => Result=F (2.0/0.0)
(EG = Very Poor) => Result=F (4.0/0.0)
 => Result=P (106.0/3.0)
```

Figure II shows the tree obtained by ADTree algorithm in WEKA Tool after using the best 13 attributes and balancing data using SMOTE.

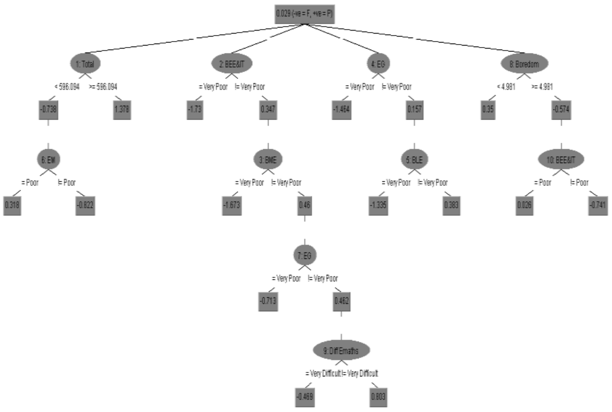

Figure 2: Tree obtained by adtreealgorithm using the best 13 attributes and data balancing

The decision tree of the Table X shows that all the students with attributes concerning marks of BLE, BME, BEE&IT, EM appear with values of very poor tend to fail. It is also shown the attribute 'Level of boredom during class' is greater than 4.981, the students tend to fail. The students who find Engineering Mathematics 'Very Difficult' tend to be in the Fail category.

Table 10. Model generated by adtree algorithm using the best 13 attributes and data balancing

```
: 0.029
| (1)Total < 596.094: -0.738
| | (6)EM = Poor: 0.318
| | (6)EM != Poor: -0.822
| (1)Total >= 596.094: 1.378
| (2)BEE&IT = Very Poor: -1.73
| (2)BEE&IT != Very Poor: 0.347
| | (3)BME = Very Poor: -1.673
| | (3)BME != Very Poor: 0.46
| | | (7)EG = Very Poor: -0.713
| | | (7)EG != Very Poor: 0.462
| | | | (9)Diff Emaths = Very Difficult: -0.469
| | | | (9)Diff Emaths != Very Difficult: 0.803
| (4)EG = Very Poor: -1.464
| (4)EG != Very Poor: 0.157
| | (5)BLE = Very Poor: -1.335
| | (5)BLE != Very Poor: 0.383
| (8)Boredom < 4.981: 0.35
| (8)Boredom >= 4.981: -0.574
| | (10)BEE&IT = Poor: 0.026
| | (10)BEE&IT != Poor: -0.741
Legend: -ve = F, +ve = P
```

Finally, it is important to note that no consensus has been detected between the previous classification algorithms about the existence of a single factor that most influences to the students' failure. However, the following set of factors (which most appear in the models obtained) can be considered as the most influential: Very Poor in BEE&IT, BLE, BME, EG ;Level of attentiveness in class less than 7.2288l; Level of Boredom during classes greater than 4.98, Level of Difficulty in Engineering Maths is 'Very Difficult'.

4. PREDICTING RESULTS: A CASE STUDY OF 60 STUDENTS

From the above discussions, it is clear that the model constructed using various algorithms can be used to predict the results of students once their details are known. The information like the scores in various subjects viz. Basic Electrical Engineering(BLE), Basic Electronics Engineering and Information Technology(BEE&IT), Basic Mechanical Engineering (BME), Engineering Graphics(EG), Engineering Mechanics(EM) and other details like level of attentiveness during classes, level of boredom during classes, their difficulty level in Engineering Mathematics etc. are known, the students result can be easily predicted using the constructed model using the algorithm ADTree. This algorithm showed maximum Accuracy, TN Rate and TP Rate when the best attributes were selected using Attribute Selection Algorithms provided in WEKA tool and the data set was balanced using SMOTE algorithm.

A case study was done with the data of 60 students out of which 50% students have opted for Computer Science and Engineering (CSE) and the remaining 50% have chosen Electronics and Communication Engineering (ECE) as their branch of study. Only the best 13 attributes were collected from the students and used for the experiment.

The experiment was done using WEKA tool. As the training set, the data set containing real life instances of 150 students with 63 attributes reduced to best 13 attributes and the data balanced using SMOTE algorithm was used. As the test file, data set about 60 students was used. The 'output prediction' option was set in WEKA tool and the experiment is done.

From the 'result lists' right click the newly obtained result and select 'visualize classifier errors' to save the prediction results.

The model predicted 37 students as Failed and 23 students as Passed. The actual results of the students were taken from the corresponding departments in college and compared with the predicted results. The actual results revealed that 36 students Failed and the remaining 24 passed. The model showed an accuracy of 91.67 % by a correctly classifying 55 instances of students. Only 5 instances were wrongly predicted by the model.

5. CONCLUSION

Predicting student failure at college can be a difficult task not only because it is a multifactor problem (in which there are a lot of personal, family, social, and economic factors that can be influential) but also because the available data are normally imbalanced. To resolve these problems, it was shown the use of different DM algorithms and approaches for predicting student failure. Several experiments were carried out using real data of first year B. Tech students in Kerala, India. Different classification approaches were applied for predicting the academic status or final student performance at the end of the course. Furthermore the study shows approaches such as selecting the best attributes and data balancing which can also be very useful for improving accuracy.

Data gathering and data pre-processing were two important tasks carried out in this work. Since quality and reliability of available information directly affects the results obtained, it was an arduous task to gather the correct information.

In general, regarding the DM approaches used and the classification result obtained, the main conclusions are as follows:

1) It was shown that classification algorithms can be used successfully in order to predict a student's academic performance and, in particular, to model the difference between Fail and Pass students.

2) The utility of feature selection techniques was also shown when a large number of attributes are involved in the study. In this case, the number of attributes used was reduced from the 67 initially available attributes to the best 13 attributes, obtaining fewer rules and conditions without losing classification performance.

3) Two different ways to address the problem of imbalanced data classification by rebalancing the data and considering different classification costs was also shown. In fact, rebalancing of the data has been able to improve the classification results obtained in TN rate, Accuracy and TP Rate.

Regarding the specific knowledge extracted from the classification models obtained, the main conclusions are as follows:

1) White box classification algorithms obtain models that can explain their predictions at a higher level of abstraction by IF-THEN rules. In this case, induction rule algorithms produce IF-THEN rules directly, and decision trees can be easily transformed into IF-THEN rules. IF-THEN rules are one of the most popular forms of knowledge representation, due to their simplicity and comprehensibility. These types of rules are easily understood and interpreted by non-expert DM users, such as instructors, and can be directly applied in decision making process.

2) Concerning the specific factor or attributes related with student failure, there are some specific values that appear most frequently in the classification models obtained. For example, the values of scores/grades that appear most frequently in the obtained classification rules and trees is the value "Very Poor"(score less than 40) in the subjects of Basic Electrical Engineering(BLE), Basic Electronics Engineering and Information Technology(BEE&IT), Basic Mechanical Engineering(BME), Engineering Graphics(EG). Other factors frequently associated with failing are being having a low level of attentiveness during classes, high level of boredom during classes and students who consider Engineering Mathematics as a difficult subject. It is also striking that the failing grades for a subject like Engineering Mathematics, that a majority of students usually pass, appear in the obtained models. A student with a good knowledge and base in Mathematics only selects Engineering as his/her higher studies option. This year of college admission has also taken students with poor background in Mathematics that the students feel this subject very difficult to study.

3) In this study, students' marks were used and did not focus solely on social, economic, and cultural attributes for two main reasons. The first is that the system obtained bad classification results when the marks were not considered. Secondly, the grades obtained by students have been previously used in a great number of other similar studies.

A case study was done with the constructed model to predict the results of 60 students with only the best 13 attributes and the model showed an accuracy of 91.67% in classifying the instances. From the models (rules and decision trees) generated by the DM algorithms, a system to alert the teacher and parents about students who are potentially at risk of failing or drop out can be implemented. As an example of possible action, let us propose that once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure.

6. FUTURE WORKS

Finally, as the next step in research, carry out more experiments using more data and also from students of different years (second, third and fourth years) to test whether the same performance results are obtained with different DM approaches. As future work, the following can be done:

1) To develop our own algorithm for classification/ prediction based on grammar using genetic programming that can be compared versus classic algorithms.

2) To predict the student failure as soon as possible. The earlier the better, in order to detect students at risk in time before it is too late.

3) To propose actions for helping students identified within the risk group. Then, to check the rate of the times it is possible to prevent the fail or dropout of that student previously detected.

7. REFERENCES

[1] C. Márquez-Vera, Cristóbal R. Morales, and S.Ventura Soto, "Predicting School Failure and Dropuout by Using Data Mining Techniques*", IEEE Journal of Latin-American Learning Technologies,* vol. 8,no. 1, February,2013.

[2] Pimpa Cheewaprakobkit, "Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in Internationa Program", Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13-15, 2013, Hong Kong.

[3] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students'Performance in Distance Education," *Knowl. Based Syst.*, vol. 23, no. 6, pp. 529–535, Aug. 2010.

[4] M. N. Quadril and N. V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques" *Global J.Comput. Sci. Technol.*, vol. 10, pp. 2–5, Feb. 2010.

[5] C. Romero and S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[6] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[7] S. Kotsiantis, "Educational Data Mining: A Case Study for Predicting Dropout—Prone Students," *Int. J. Know. Eng. Soft Data Paradigms*,vol.1, no. 2, pp. 101–111, 2009.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Synthetic Minority Over-Sampling Technique," *J. Artif. Intell. Res.*,vol. 16, pp. 321–357, Jun. 2002.

[9] J. Cendrowska, "PRISM: An Algorithm for Inducing Modular Rules," *Int.J. Man-Mach. Stud.*, vol. 27, no. 4, pp. 349–370, 1987.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufman, 1993.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York, USA: Chapman & Hall, 1984.

[12] Y. Freund and L. Mason, "The Alternating Decision Tree Algorithm,"*Proc. 16th Int. Conf. Mach. Learn.*, 1999, pp. 124–133.

[13] Sunita B Aher,Lobo L.M.R.J "Data Mining in Educational System using WEKA", *International Conference on Emerging Technology Trends (ICETT)* 2011 Proceedings published by International Journal of Computer Applications® (IJCA) 20.

[14] Suchita Borkar, K. Rajeswari," Predicting Students Academic Performance Using Education Data Mining", *IJCSMC,* Vol. 2, Issue. 7 July 2013, pg.273 – 279.

[15] R.S.J.D Baker and K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions" , Journal of Educational Data Mining, 1, Vol 1, No 1, 2009.

[16] I.H. M. Paris, L.S. Affecndy and N.Musthafa, "Improving Performance Prediction using Voting technique in data Mining", World Academy of Science, Engineering and Technology World Academy of Science, Engineering and Technology, Vol 38,2010.

[17] T.Nghe, J.Paul , Aneek and Peter Heddawy, "A Comparitive Analysis of Techniques for Predicting Academic Performance", Paper presented at 37th ASEE/IEEE Conference, Frontiers in Education Conference – Global Engineering: Knowledge Without Borders, OpportunitiesWithout Passports, Milwaukee,WI,October 10-13,2007.

[18] P.Cheewaprakobkit, "Study of Factor Analysis Affecting Achievements of Undergraduate", Paper presented at International Multi Conference of Engineers and ComputerScientists, IMECS , Hong Kong, HK, March 13 - 15, 2013.

[19] B. Sen, E. Uçar and D. Delen, "Predicting and Analyzing Secondary Education Placement-Test Scores: A Data Mining Approach", Expert System with Application, Volume 39, Issue 10, 2012.

[20] M. Wook, Y.H.Yahaya, N. Wahab, M. R.M. Isa, N. F. Awang aInternational Conference nd H.Y. Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques, Paper presented at International Conference of Computer and Electrical Engineering, ICCEE. December 28-30. 2009.

[21] B. M. Bidgoli, D.Koshy, G.Kortemeyer, W.F.Punch, "Predicting Student Performance: An Applicant of Data Mining methods with an educational web based system" , 33rd ASEE/ IEEE .frontiers in Education Conference 20004.

[22] E. Osmanbegovic and M. Suljic, " Data mining Approach for Prediction of Student Performance" Economic Review - Journal of Economics & Business Vol. 10, issue 1, 2012.

[23] M. Ramaswami, and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science, Vol. 7, Issue 1, No. 1.of 2010.

[24] N. S. Shah, "Predicting Factors that Affect Students ' Academic Performance By Using Data Mining," Pakistan Business Review, January 2012.

[25] D.Kabakchieva, "Predicting Student Performance by using DataMining methods for classification.", Cybernetics and Information Technologies, Volume 13,2013.

[26] R. R. Kabra, R.R, Bichkar ," Performance Prediction of Engineering Students using Decision Trees", International Journal of Computer Applications, Volume 36, No.11, 2011.

[27] Z. J .Kovacic, "Mining Students Enrolment Data", Paper presented at Proceedings of Informing Science & IT Education Conference (InSITE) ,Casinio Italia, June, 19-24,2010.